DOCUMENT RESUME

ED 255 537 TM 850 037

AUTHOR Jones, Douglas H.; And Others
TITLE Estimating Ability with Three Item Response Models when the Models are Wrong and Their Parameters are Inaccurate.
INSTITUTION Educational Testing Service, Princeton, NJ. Program Statistics Research Project.
SPONS AGENCY Air Force Human Resources Lab., Brooks AFB, Texas.
REPORT NO ETS-PSR-TR-84-46; ETS-RR-84-26
PUB DATE Jul 84
CONTRACT F41689-82-C-0020
NOTE 59p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Ability; Aptitude Tests; *Error of Measurement; Estimation (Mathematics); *Goodness of Fit; *Latent Trait Theory; *Mathematical Models; *Maximum Likelihood Statistics; Predictive Measurement; Statistical Bias; Testing Problems
IDENTIFIERS Armed Services Vocational Aptitude Battery; Item Characteristic Curves

ABSTRACT

How accurately ability is estimated when the test model does not fit the data is considered. To address this question, this study investigated the accuracy of the maximum likelihood estimator of ability for the one-, two- and three-parameter logistic (PL) models. The models were fitted into generated item characteristic curves derived from the Armed Services Vocational Aptitude Battery. Results indicated that the 3-PL model is the least biased, but that for lower abilities it had both the highest mean square error and was the most fragile with respect to sampling fluctuations of the parameter estimates. It is recommended that the current practice of utilizing some method of restricting the variability of parameters continue to be employed (such as Bayesian priors, robust estimation, or fixing some parameters to be constant). (Author/BS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H. Weidenmiller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Estimating Ability With Three Item Response Models When the Models are Wrong and Their Parameters are Inaccurate

by

Douglas H. Jones*
Howard Wainer
Bruce Kaplan

Program Statistics Research
Technical Report No. 84-46

Research Report RR-84-26

July 1984

*Dr. Douglas H. Jones is affiliated with Rutgers University and the Advanced Statistical Technologies Corporation

Copyright © 1984 by Educational Testing Service. All rights reserved.

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

Reproduction of any portion of a Program Statistics Research Technical Report requires the written consent of the author(s).

Table of Contents

		Page
I.	Introduction	2
II.	The Structure of the Study	3
III.	What About Longer Tests?	13
IV.	Discussion and Implications	14
V.	References	16
VI.	Mathematical Appendix	17

Abstract

We investigate the accuracy of the maximum likelihood estimator of ability for the one- two- and three-parameter logistic models. This is accomplished by fitting of those models to generated item characteristic curves derived from the Armed Services Vocational Aptitude Battery. The mathematical details as to how this is accomplished and our results are included. Among these results is the finding that the 3-PL model is the least biased, but that for lower abilities it has both the highest M.S.E. and is the most fragile with respect to sampling fluctuations of the parameter estimates. It is recommended that the current practice of utilizing some method of restricting the variability of parameters continue to be employed (i.e., Bayesian priors, robust estimation, fixing some parameters to be constant).

Acknowledgements

This research was supported in part by the Program Statistics Research Project of the Educational Testing Service and by contract number F41689-82-C-0010, from the Air Force Human Resources Laboratory to Educational Testing Service, Howard Wainer, Principal Investigator.

We would like to thank Isaac Bejar, James Earles, Frederic M. Lord, Malcolm Ree, Paul Rosenbaum, and David Thissen for helpful conversations.

I. INTRODUCTION

In this investigation we are principally concerned with one question: How accurately is ability estimated when the test model does not fit the data? There are a variety of sub-issues in this, among them:

a) what sort of estimator?

b) what sort of model?

c) how bad was the fit?

d) how close is the actual error to the model-based standard error?

e) can we improve things? how?

To place these questions into context, let us consider a testing situation that will become increasingly commonplace in the future. Specifically, that of an individual who sits down at a desk or computer terminal and takes a test designed to measure single or multiple abilities. The statistical mechanism behind the test scoring and administration will almost surely be based on an Item Response Theory (IRT). The person's score can thus be reported as an ability parameter in that IRT model (or parameters in a multiple ability case).

Standards of good practice require that along with the estimate of the examinee's ability must be produced a measure of the accuracy of that estimate. This measure has several components:

i) the variability reported by the model--this is often portrayed as either Fisherian information or its inverse, obtained from the likelihood equations.

ii) the variability introduced through the lack of perfect fit of the model to the data.

iii) the variability due to variations in human performance from one time to another.

iv) the variability in item parameter estimates that one expects from sampling fluctuations.

To examine the effects of these possible sources of variation we designed a reasonably complex study. This study includes all of the above sources except (iii). In the following section we shall describe the structure of the study, the logic that generated that structure, and the results we obtained. In the appendix we provide the mathematical details that are required for the reader to replicate our findings.

II. THE STRUCTURE OF THE STUDY

a. Outline

1) Choose a "true model" with known parameters.

2) Fit the one-, two-, and three-parameter logistic functions to the ICC's generated by the true model.

3) Explore the response surface of the m.s.e. of $\hat{\theta}$ in the region of the estimated parameters of the models being considered. Examining:

(i) How well we might do (minimum m.s.e.)

(ii) How badly we might do (maximum m.s.e.)

(iii) The most likely outcome (expected m.s.e)

b. The 'True' Model

The choice of the model that is assumed for the data structure is critical. If we choose one that is too simplistic (i.e. conforming exactly to the Rasch Model), the results of our study will have uncertain generalizability -- that is give too optimistic results. If we choose a model that is too

pathological, the pessimistic results might poin. us toward an operational model that is too cumbersome. As a result of considerations like these, and with a desire to make our results as directly relevant to current practice as possible, we decided to fit actual data from a large scale administration of a Armed Services Vocational Aptitude Battery (ASVAB) with a very general IRT model originally formulated by Bock (1972) and subsequently modified by Samejima (1979), Bock (Personal Communication) and Thissen and Steinberg (1983). This model, the so-called "BSBT" model is described in detail elsewhere (Wainer and Thissen, 1984) and will not be repeated here. It is sufficient for the purposes of this paper to know that it is quite general, but its complex 14 parameter structure makes it too cumbersome for many situations where an IRT approach might otherwise prove fruitful. After fitting ASVAB data with this model by using the methods of Thissen and Steinberg (1983), we chose four items whose Item Characteristic Curves (ICCs) were well defined by the data set, and which seemed to best represent the types of curves found on the ASVAB. These four ICC's are shown in Figure 1.

Insert Figure 1 About Here

Note that item 3 would be fit well by a model with a zero lower asymptote, whereas item 2 requires a non-zero asymptote. Of some importance is the up-turn at the low end of items 1 and 4 which none of the common IRT models will fit.

The use of this model and a four item test can thus easily provide us with the true probabilities of getting each item correct for any specified ability. We shall be concerned with ability at five different levels. Shown in Table 1 are the true probabilities of getting each item correct for these

five ability levels. These will be used subsequently in the study of the variability of the ability estimates obtained with the models tested.

Insert Table 1 About Here

c. Fitting the Three Models

The three most commonly used IRT models were then fit to the ICCs shown in figure 1. This was done simply as a binary regression in which 100 "abilities" were drawn from a distribution that was assumed to be N(0,1) and is thus analogous to conditional maximum likelihood estimation given the ability parameters. The values of the estimated parameters for each of the three models is shown in Table 2 along with the model generated estimates of the standard errors of these parameters.

Insert Table 2 About Here

In figures 2, 3 and 4 are shown the estimated ICC's for each of the three models; the 'true ICC' is included for purposes of visual comparison. While we invite the reader to investigate these

Insert Figures 2, 3 and 4 About Here

curves, a few features are worthy of note. Specifically, the 1-PL tends to make all the items seem easier than the other models because it interprets the high guessing rate as a lot of individuals of low ability getting items right. The 2-PL accommodates these non-zero lower asymptotes by interpreting them as evidence of a gradual slope (low discrimination). The 3-PL allocates the various pieces of the curve in a way that seems to make sense in casual

observation, but still the misfit at the low end of the items 1 and 4 is a problem — the impact of which we will assess shortly. It also provides standard errors of the discrimination parameter that are much larger than for the other two models.

The three models tested here — the 1-PL, 2-PL and 3-PL — are in some sense straw-men. Current usage of item response models generally does not use these models in their pure form. Most computer programs that advertise maximum likelihood estimation of the parameters of these models use some kind of restrictive bounds on the value of the parameters. This can take the form of rough boundaries that will not permit a parameter to go beyond (i.e. restrict 'c' to the interval [0,1]; or take a formal Bayesian approach in which the inclusion of a prior distribution on the parameter prevents it from straying far afield.

Thus, the model fitted by, say, the program LOGIST (Wood & Lord, 1976) is not what we call the 3-PL. It is probably more aptly called the 2.2-PL. We shall go into this more deeply in the discussion section, but we felt it is important to point out that the horrible results which we are about to describe apply to these models when they are fitted with an unadulterated maximum likelihood method. These results point out why sensible investigators do not use the pure form of the m.l.e..

d. The Exploration of the Response Surfaces of the mse ($\hat{\theta}$)

The material in sections b and c provide the background for the actual study. In this section we will provide both a description of what we did and a rationale for doing it.

Using the true model and a specified value of the true ability parameter, θ, we can generate the distribution of response patterns and their relative

frequency. There are 2^4 possible patterns, and for any given value of θ the relative frequency of those patterns varies in a calculable way. Going backwards, if we collect a large enough number of these response patterns we can estimate the value of θ that is most likely to have generated them. The variation that we observe in this experiment can be characterized by some error statistic. A common one, which we shall adopt here, is the mean square error (mse). This statistic can be decomposed into two parts: the bias (the difference between the expected value of $\hat{\theta}$ and the true value) and the variance (the stochastically generated variation around the expected value). We shall look at these two parts separately, and in combination. The estimator $\hat{\theta}$ does not have the required number of finite moments for mse to be meaningful, however we approximate $\hat{\theta}$ by a stochastic differential having the necessary moments, see equation (16) of the Appendix.

The mse in ability when the model is assumed correct is what is usually conveyed in the standard error generated by the model. What is this statistic when the hypothesized model is not correct? We wanted to examine this question from two directions. First the simple task of generating the mse ($\hat{\theta}$) for a given hypothesized (but wrong) model was accomplished by generating a set of response patterns under the true model with known θ, but interpreting these patterns as coming from a different model (the 1, 2 or 3-PL) with parameters as estimated. The results of this are shown in Table 3. This table contains the essence of this investigation. Note carefully that the entries in this table must be divided by test length (n) to obtain estimates of the m.s.e. for any individual situation. Thus, the actual values of the mean square error will not be as large as those tabled.

Insert Table 3 About Here

As we can see, for this particular point estimate of the item parameters for each of the models several outcomes are immediately evident. First, the 3-PL is by far the least biased, and the 1 and 2-PL models have quite serious bias troubles at the low end (θ = -2, -1, and 0). In fact, both models do not maintain the correct order of abilities in these ranges. Note that these are asymptotic results — larger samples would not help. Moreover, because these are asymptotic the estimates of θ are always finite.

The second result is that the variance (see equation 16 of the mathematical appendix) of the 3-PL is quite large at the low end — so large in fact that it wipes out the advantage that this model possessed over the two others when one considers mse at the low end of the ability distribution. At higher abilities the 3-PL does very well indeed.

Looking at the mse(θ) at only one point in the item parameter space is interesting, but is not as informative as examining the entire space of the sampling distribution of item parameters. We can expect reasonably profound differences from the results in Table 3 in this exploration. The primary reason for this is that although the 1-PL is rather seriously biased at some ability levels, the sampling distribution of its parameters is relatively tightly constrained compared to that of, say, the 3-PL. This was shown by Lord (1980) and Thissen and Wainer (1982).

Thus, the next step in our study is explore the space of the sampling distributions of the item parameters for the three models of interest. We do this by sampling points within the 95% highest density region of the parameter

space for this four-item test (4 space for the 1-PL, 8 space for the 2-PL and 12 space for the 3-PL) and keeping track of the values of the mse and its two components as we go. We then record:

a) the best we could do (minimum mse in the region)

b) the worst we might do (maximum mse in the region)

c) the expected mse (integrating over the region, weighting by the density.

Let us examine (c) first. It is helpful to make the distinction between the conditional m.s.e. shown in table 3 — conditioned on a specific value of the item parameters — and the unconditional m.s.e., obtained by intergrating over the entire region around the point estimate of the item parameters. If we find that there is very little difference between the conditioned and unconditioned estimates, it would imply that the anticipated sampling fluctuations in parameter estimates would have little effect on the accuracy of our estimates. If, on the other hand there are substantial differences, it tells us that the estimate of the m.s.e. is quite dependent upon the precision of the estimates, and that this fragility of estimation accuracy could be damaging. The results of this comparison are shown in Table 4.

Insert Table 4 About Here

It is clear that at low ability levels the fragility of the 3-PL could be troublesome. Another way of characterizing this fragility is to construct an index which compares the unconditioned m.s.e. with the conditioned m.s.e. This index will be denoted L and is:

$$L = (\text{Unconditioned mse} - \text{conditioned mse})/\text{unconditioned mse}$$

$$= (\text{mse}(\hat{\theta}) - \text{mse}(\hat{\theta} \mid a_o, b_o, c_o)) \,/\, \text{mse}(\hat{\theta})$$

Note that this can also be written:

$$L = \frac{(\text{conditioned mse})^{-1} - (\text{unconditioned mse})^{-1}}{(\text{conditioned mse})^{-1}}$$

which makes clearer an alternative interpretation as relative information loss.

A plot of relative information loss against ability provides insights into the performance of the three models (see Figure 5). We note that the

---

Insert Figure 5 About Here

3-PL has the highest loss (especially at low ability levels) when moving from the conditional to the unconditional estimate. The 1-PL is relatively constant and low with regard to information loss and the 2-PL model lies intermediate between them. This does not mean that the 1-PL is terrific, rather it means that its performance curve is quite flat in the region of the m.l.e. The 3-PL, on the other hand, varies a great deal. Thus, one might be fooled into thinking that one had done a lot better than was in fact the case. As we shall see shortly, the uniform performance of the 1-PL is, in fact, uniformly mediocre.

Before going on to the further exploration of the m.s.e. surface let us pause briefly to summarize what we have so far discovered:

1) The 3-PL has the least bias of any of the three models studies -- by a long shot.

2) Despite this small bias, the 3-PL has the largest m.s.e. of any of the models at low ability levels.

3) It follows from (1) and (2) that the proportion of bias in the m.s.e. for the 3-PL is much lower than for the other two models. This is shown in Figure 6.

4) The gigantic size of the m.s.e. for the 3-PL at low ability levels may be a bit misleading. If instead we look at its reciprocal -- which would be an information measure -- we note that although the 3-PL has

---

Insert Figures 6 and 7 About Here

---

the lowest information of the three models at low ability levels, the other two models are nothing to write home about. At higher ability levels (+1 and +2) the 3-PL does quite well (see Figure 7).

5) The results in (1) through (4) deal with the conditional estimates of θ at the point a_0, b_0, c_0. When we consider the fragility of these estimates to plausible fluctuations in the region around that point (the unconditioned estimates) we find that the 3-PL's superiority at high ability levels might be more apparent than real in a practical application (see Figure 5). We shall investigate this more fully next.

The fragility of the estimates of the m.s.e.($\hat{\theta}$) that was suggested in the comparisons between the conditioned and unconditioned estimates becomes clearer still when we examine the broad range of values possible in the 95% highest density region around m.l.e. of the item parameters. The minimum, 25th %ile, 75%ile and the maximum values for the m.s.e. for each of the three models are shown in Table 5.

Insert Table 5 About Here

The same structure that was suggested earlier continues to hold — that the 1-PL [illegible] if not accurate, whereas the 3-PL can be quite accurate, but can blow up horribly as well. The most surprising result is that the conditioned estimate (the estimate of θ based upon the m.l.e. of the item parameters) is <u>not</u> the value which minimizes the m.s.e.. The fact that this is so is not the surprising aspect, for we know that the m.l.e. would only be optimal asymptotically when the model fits. In this situation none of the models fit perfectly. What is surprising is how badly we can do. For example, consider at $\theta = -2$ the 3-PL conditional estimate has an m.s.e. of 619, but that within a plausible region it could be as low as 19. This same structure does not seem to occur for the 1-PL nor 2-PL, in that the conditional estimate seems closer to the minimum for these models.

The results in Table 5 can be examined in another way, as relative efficiencies. In this instance we will follow the usual definition and denote

$$\text{relative efficiency } (\theta) = \text{Minimum m.s.e.}(\theta_i) \;/\; \text{m.s.e.}(\theta_i)$$

where this minimum is taken over all models at a given ability. Thus, at each ability level some estimator will have 100% efficiency. The results of this

are shown in Table 6. The results are quite straightforward — at high ability levels (+1 and +2) the 3-PL dominates; although even in this situation we see that there is great variability in the region of the conditional solution, and that substantial improvement is possible over the conditional estimate.

Insert Table 6 About Here

At lower ability levels the choice is not so clear. The 1-PL gives stable but unspectacular results whereas the 2-PL provides the possiblity of the best and the worst performance.

III. WHAT ABOUT LONGER TESTS?

So far we have constrained ourselves to exploring the mse for a four-item test. The principal reason for this is the computational effort involved in integrating over a space whose dimensionality increases exponentially with the length of the test. However, we can evaluate the mse for the conditional estimate at several test lengths of interest, expanding the test by including new items with identical characteristics of the four used so far. This corresponds to expanding Table 3. This expansion is shown in Table 7 for test lengths of 4, 16, 48 and 96 items. We choose these lengths since:

48 and 96 emmulates conventional size tests

16 emmulates an adaptive test

4 emmulates the initial part of an adaptive test in which a preliminary estimate of ability is obtained.

Insert Table 7 About Here

Examining these results reinforces the results seen in Table 3. We note that as test length increases the proportion of the mse contributed by the bias diminishes. The 3-PL dominates when ability is greater than 0; but for lower ability levels the variance of the unconstrained 3-PL is still overwhelming its lower bias.

IV. DISCUSSION AND IMPLICATIONS

The results of this study must be understood within the proper context. We have explored the fitting of some reasonable data with three models and examined how the m.l.e. for these models fluctuates. The model we have designated 1-PL is actually quite close to the Rasch model that is used in practice, and so the results can be generalized. The 2-PL is also not too far from the model used in practice, although extreme estimates of ability are commonly avoided through the use of Bayesian priors. This would avoid the drastic fluctuations in performance observed. The 3-PL we have tested does not resemble the version generally used. The erratic performance we have noted has been manifested in other ways in various applications of this general model and steps are usually taken to prevent it. These precautions involve the limiting of the variation of the item parameters so that the standard errors that we used to characterize their m.l.e. would be vast overestimates. In addition, priors are sometimes used on the ability parameter to assure that its estimate is finite. Thus, the 3-PL results we have provided should not be thought of as characterizing the way this model would behave when properly controlled. Our point in including this version was to fully illustrate just how important these restraints are in obtaining usable results.

This brings us, finally, to the essense of our study. No one model dominates the others, although under favorable conditions all can do a tolerably good job. That we were not able to accurately estimate ability in the lower regions is no surprise — how can one obtain good estimates when the examinees in that region are providing more noise than signal? The best we should hope for in this instance is that our measurement models yield a plausible and reliable measure of our own ignorance. Models without a guessing parameter are quite biased in this area (-2 to 0) but are pretty stable. The 3-PL reduces this bias but at a terrible expense in variance. The obvious solution (one that is taken by LOGIST, for example) is to use a lower asymptote but not to estimate it. The non-monotonicities observed in some of these ASVAB items did not prove to be a serious hinderence.

Thus, our final conclusion that while we have not shown exactly what one must do to obtain useful estimates — i.e. bound parameters, use Bayesian priors, use robust estimators — we feel that we have demonstrated that one must do something; the naive use of maximum likelihood estimation with any of the models examined is sure to provide unhappy results.

V. REFERENCES

Apostol, T.M. (1957) Mathematical Analysis. Reading, MA: Addison-Wesley.

Bock, R.D. (1972) Estimating item parameters and latent abilities when the responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.

Huber, P.J. (1967) The behavior of maximum likelihood estimates under nonstandard conditions, in: Proceedings Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, University of California Press, Berkeley.

Jones, D.H. (1982) Redescending M-Type Estimators of latent ability. Princeton, NJ: ETS Program Statistics Research Technical Report No. 82-30 and Research Report No. 82-15.

Lord, F.M. (1980) Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Samejima, F. (1979) A new family of models for the multiple choice item. Research Report #79-4, Department of Psychology, University of Tennessee.

Serfling, R.J. (1980) Approximation Theorems of Mathematical Statistics. New York: John Wiley and Sons.

Thissen, D. and Wainer, H. (1982) Some Standard Errors in Item Response Theory, Psychometrika, 47, 397-412.

Thissen, D. and Steinberg, L. (1983) A model for multiple choice items. (Submitted to Psychometrika).

Wainer, H. and Thissen, D. (1984) Estimating ability with the wrong model (manuscript in preparation).

Wood, R.L. and Lord, F.M. (1976) A user's guide to LOGIST. Research Memorandum 76-4. Princeton, NJ: Educational Testing Service.

VI. MATHEMATICAL APPENDIX

a) Definitions of the One, Two and Three Parameter Item Response Functions

First, let us define the maximum likelihood estimator of ability for the one, two, and three-parameter logistic response models conditioned on the values of the item characteristic parameters. Let u_i denote the binary response to the i-th item and the probability of correct response given the ability θ, a scaler by

$$(1) \qquad Pr(u_i = 1|\theta) = P_i(\theta) \quad i=1,2,\ldots,n$$

where n denotes the total number of items in the test. Let a_i, b_i and c_i denote the characteristic parameters for the i-th item and each of the three logistic response functions as follows:

one-parameter logistic (1-PL):

$$(2) \qquad P_i(\theta) = \{1 + \exp[-a(\theta-b_i)]\}^{-1} \quad i=1,2,\ldots,n$$

where $a>0$

two-parameter logistic (2-PL):

$$(3) \qquad P_i(\theta) = \{1 + \exp[-a_i(\theta-b_i)]\}^{-1} \quad i=1,2,\ldots,n$$

where $a_i>0$

three-parameter logistic (3-PL):

$$(4) \qquad P_i(\theta) = c_i + (1-c_i)R_i(\theta) \quad i=1,2,\ldots,n$$

where $0<c_j<1$ and $R_i(\theta)$ is a 2-PL.

The three models are hierarchical, being obtained from the 3-PL by setting some of the item parameters equal to zero and or equal to constants as the

index i varies. Since there are many item parameters for the entire test we can simplify the notation by denoting all the item parameters by the matrix Ψ_n,

$$\Psi_n = \begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ a_n & b_n & c_n \end{bmatrix} . \tag{5}$$

For a given matrix of item parameters, the induced likelihood of a response sequence $u_1,u_2,\ldots,u_n$ given ability θ is, assuming local independence,

$$L(\theta) = L(\theta;\, u_1,\ldots,u_n,\Psi_n) = \prod_{i=1}^{n} P_i(\theta)^{u_i} Q_i(\theta)^{1-u_i} \tag{6}$$

where $Q_i(\theta) = 1-P_i(\theta)$. The derivative of the logarithm of $L(\theta)$ produces the likelihood equation defining the maximum likelihood estimate $\hat{\theta}_n$:

$$\hat{\theta}_n \text{ solution to } \sum_{i=1}^{n} w_i(\theta)[u_i-P_i(\theta)] = 0 \tag{7}$$

where $w_i(\theta) = d/d\theta \log[P_i(\theta)/Q_i(\theta)] = [d/d\theta P_i(\theta)][P_i(\theta)Q_i(\theta)]^{-1}$.

Now we characterize the asymptotic behavior of the maximum likelihood estimator of ability as the number of items, n, becomes large. For now we assume that the matrix of item parameters Ψ_n,n=1,2,..., are nonstochastic and given for each logistic response model and for each n.

b) The Construction of the Expected Loss

The primary contribution of the mathematical portion of this paper is the derivation of properties of the mle of ability (7) when the true likelihood

of responses is different from the induced likelihood (6). The mathematical formulation of these properties is needed to compare estimators when the true likelihood is generated by the Bock-Samejima model. More precisely, the formulation is needed for expansions of the comparison criteria which enable computational feasibility.

Ideally, we would like to compare the mean-square errors of estimators (MSE's) of a parameter θ^*, say. (We leave for later the case where the estimators do not have a sufficient number of finite moments for the mean-squared error to be meaningful.) It turns out that the particular expansion of the MSE needed for computation depends on whether the induced and true likelihoods agree.

When the true and induced likelihoods agree, we expect the estimators to be consistent; that is, they converge to θ^* as the sample size becomes infinite. Thus, letting $\mu_n = E\hat{\theta}_n$, the computational expansion for the MSE that would be useful is

$$E(\hat{\theta}_n - \theta^*)^2 = E(\hat{\theta}_n - \mu_n)^2 + (\hat{\mu}_n - \theta^*)^2 \quad . \tag{8}$$

Expansions of the estimator $\hat{\theta}_n$ then would be used to estimate the first and second terms. Since the estimators are consistent it would be useful to determine the exact order of $(\mu_n - \theta^*)^2$. For an unbiased estimator this term is exactly zero; for most estimators the order is n^{-2} (because $\mu_n - \theta^*$ is $O(n^{-1})$). For the typical situation, $\hat{\theta}_n$ is asymptotically normal (AN) for norming constants μ_n and $\sigma/\sqrt{n}$, and thus the second term is included to improve the small sample accuracy of the approximation to (8).

When the true and induced likelihoods disagree, an estimator is not consistent but typically converges to some scaler, θ_0, other than θ^*,

considered to be the true parameter. The second term of (8) now becomes more important since it is no longer asymptotically negligible. The needed expansion of the MSE is

$$(9)\quad E(\hat{\theta}_n/\theta^*)^2 = E(\hat{\theta}_n-\mu_n)^2 + (\mu_n-\theta_0)^2 + 2(\mu_n-\theta_0)(\theta_0-\theta^*) + (\theta_0-\theta^*)^2 .$$

Typically, $\mu_n \neq \theta_0$ but μ_n converges to θ_0 and the orders of the two middle terms need to be determined. Clearly, the second term is of smaller order than the third. In the typical situation, $\hat{\theta}_n$ is $AN(\mu_n, \sigma\sqrt{n})$ and the second term is negligible, so approximations to (9) are the same yielded by

$$(10)\quad E(\hat{\theta}_n-\theta^*)^2 = E(\hat{\theta}_n-\mu_n)^2 + 2(\mu_n-\theta_0)(\theta_0-\theta^*) + (\theta_0-\theta^*)2$$

where the middle term is not negligible but is of the same order as the first.

Typically, estimators derived from two different induced likelihoods will converge to different limits. When this happens, the last term in equation (10) determines the relative performance of the two types of estimators. In this case, we ignore the middle term on the right-hand side of (10) since we want to know the reason why one estimator out-performs another in simple and familiar terms. For example, when one type of estimator is superior, is it because its limit θ_0 is close to θ^*, as reflected by $(\theta_0-\theta^*)^2$, or is it because it is very efficient in estimating θ_0, as reflected by $E(\hat{\theta}_n-\mu_n)^2$?

Thus, we will use a different notion of mean-squared error than the traditional mean-squared error, defined as

$$(11)\quad MSE = E(\hat{\theta}_n-\mu_n)^2 + (\theta_0-\theta^*)^2 .$$

MSE (11) has two components of expected loss: the first due to the inefficiency of the estimator, reducible by increasing sample size; and the second due to bad choice of item response model induced likelihood. We refer to these two terms in order as the variance of $\hat{\theta}_n$ and the bias squared of the model.

While it is obvious that increased number of items results in increased costs, if we assume that it costs more to reduce the bias of a model by increasing its complexity, then MSE (11) is a natural cost function for item response theory. As the main part of this paper illustrates, the more complex a model is, the higher the variance of its mle. Thus, the gain in reducing bias of the model may have to be balanced with the loss of efficiency.

MSE (11) is not dominated by $E(\hat{\theta}_n-\theta^*)^2$. It is clear from (10) that MSE (11) is at most $E(\hat{\theta}_n-\theta^*)^2$ whenever $\mu_n-\theta_0$ and $\theta_0-\theta^*$ are of the same sign. Thus, comparisons between estimators based on MSE (11) may be qualitatively different from those based on (10).

c) The True Parameter, θ^*

In order for the MSE (11) to be meaningful we assume that the sequence of theoretical expected values of item responses, $u_i, i=1,2,\ldots$, satisfy an unidimensional model, possibly different from (2), (3) or (4). That is, assume that for each respondent, there exists a scaler, θ^*, and a function P_i^* on the real line such that

$$0 \leq P_i^*(\theta) \leq 1 \text{ for } -\infty < \theta < \infty \tag{12a}$$

$$E(u_i|\theta^*) = P_i^*(\theta^*) . \tag{12b}$$

d) Regularity Conditions on Item Sequences Comprising Tests

Surprisingly, under mild regularity conditions on $\{P_i : i=1,2,\ldots\}$, it is possible to fully stipulate equations for computing the limit θ_0 and variance of an estimator sequence $\{\hat{\theta}_n\}$. We will assume regularity conditions that are more restrictive than need be, however, they allow an easily understandable derivation of the required computational formulas.

Assume that there are a finite number of generic item types with true item response functions $\{\Pi_k^* : k=1,2,\ldots,k\}$. That is, if a particular item i is of generic type k, then a respondent with true ability θ^* has $E(u_i|\theta^*) = \Pi_k^*(\theta^*)$.

Likewise, assume that each family of models (2), (3) and (4) have generic item response functions $\{\Pi_k : k=1,2,\ldots,k\}$ corresponding to each item type. We assume that these conditions hold in the usual condition when all items are unique.

Define π_{kn} to be the proportion of items in a test of n items that are of type k. Assume that these proportions converge as n goes to infinity and define π_k to be the limit of π_{kn}. For example, if a test of size n has k=n items and larger tests of size 2n, 3n etc., are constructed by replicating the k unique items, then $\pi_{kn}=\pi_k=K^{-1}$.

e) The Limit θ_0 of an Estimator Sequence

Under the regularity conditions it follows from the strong law of large numbers that for each of the item response models (2), (3) or (4), the derivative of the log-likelihood (7) is

$$(13) \qquad n^{-1}\sum_{i=1}^{n} w_i(\theta)[u_i - P_i(\theta)] \to \sum_{k=1}^{k} w_k(\theta)[\Pi_k^*(\theta^*) - \Pi_k(\theta)]\Pi_k$$

where the w_k of the right-hand side is $w_k(\theta) = [(d/d\theta)\Pi_k(\theta)]\{\Pi_k(\theta)[1-\Pi_k(\theta)]\}^{-1}$.

Define the function $H(\theta)$ as

$$(14) \qquad H(\theta) = \sum_{k=1}^{k} w_k(\theta)[\Pi_k^*(\theta^*) - \Pi_k(\theta)]\Pi_k \ .$$

and θ_0 as the solution to the equation

$$(15) \qquad H(\theta) = 0 \ .$$

It follows from (13) that $\hat{\theta}_n$ converges to θ_0 with probability one.

2) The Approximation to $E(\hat{\theta}_n - \nu_n)^2$

The variance $E(\hat{\theta}_n - \nu_n)^2$ does not exist for any mle of the response models (2), (3) or (4). The property of an mle of (2), (3) or (4) that causes this unfortunate happenstance is simple to characterize. If item responses are all correct or are all incorrect, then $\hat{\theta}_n = \infty$ or $\hat{\theta}_n = -\infty$ respectively.

We may deal with this situation in one of two ways. The first way is to modify the squared error loss in the first term of MSE (11) to take on a predetermined finite value $M > 0$ whenever the event $|\hat{\theta}_n| = \infty$ occurs. The probability of this event is derivable using the true likelihood. Thus, using the theorem of conditional probabilities, it is possible to expand MSE (11) into terms that are finite. The contribution $M \cdot P\{|\hat{\theta}_n| = \infty\}$ converges to zero exponentially fast thus making it negligible with respect to $(\theta_0 - \theta^*)^2$. So MSE (11) could be approximated by taking all expected values conditional on the event $|\hat{\theta}_n| < \infty$. This overcomes the problems caused by the event $|\hat{\theta}_n| = \infty$.

The second way is to make use of the expansion for $\hat{\theta}_n$ that one would use to approximate the variance conditional on $|\hat{\theta}_n|<\infty$. That is we would approximate $\hat{\theta}_n$ by a statistic that has the needed moments. We will see shortly that this approximation is akin to the Newton-Rapheson algorithm for solving (7).

The required approximation to $\hat{\theta}_n$ will be derived later. For now, we present its formula and derive its variance. The approximation is

$$(16)\quad \hat{\theta}_n \doteq \theta_0 + n^{-1}\sum_{i=1}^{n}\theta_1(u_i;\theta_0) + \frac{n^{-2}}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\theta_2(u_i,u_j;\tilde{\theta}),$$

where $|\tilde{\theta}-\theta_0|<|\hat{\theta}_n-\theta_0|$. Letting $h(\theta,u)=w_1(\theta)[u_i-P_1(\theta)]$ and $H_1(\theta)=(\partial/\partial\theta)H(\theta)$ where $H(\theta)$ is defined in (14), the terms in (16) are defined as

$$(17)\quad \theta_1(u_i;\theta_0) = -h(\theta_0,u_i)/H_1(\theta_0)$$

$$(18)\quad \theta_2(u_i,u_j;\theta) = (\partial/\partial\theta)[h(\theta,u_i)h(\theta,u_j)/H_1(\theta)]/H_1(\theta_0)$$

It is clear that approximation (16) has finite second moments. The determination of the moments require the first and second moments of $n^{-1}\sum\theta_1(u_i;\theta_0)$. If we assume n is a multiple of k and $\Pi_{ik}=\Pi_k$ for each n, then

$$(19)\quad En^{-1}\sum_{i=1}^{n}\theta_1(u_i;\theta_0) = -n^{-1}\sum_{i=1}^{n}w_i(\theta_0)[Eu_i-P_i(\theta_0)]/H_1(\theta_0)$$

$$= -n^{-1}\sum_{i=1}^{n}w_i(\theta_0)[P_i^*(\theta^*)-P_i(\theta_0)]/H_1(\theta_0)$$

$$= -\sum_{k=1}^{k}w_k(\theta_0)[\Pi_k^*(\theta^*)-\Pi_k(\theta_0)]\Pi_k/H_1(\theta_0)$$

$$= -H(\theta_0)/H_1(\theta_0)$$

$$= 0$$

Using (18) and $H(\theta_0)=0$, it can be shown that the last term of (16) is $O_p(n^{-1})$.

$$(20) \qquad \hat{\theta}_n - \mu_n \doteq n^{-1} \sum_{i=1}^{n} \vartheta_1(u_i;\theta_0) + O_p(n^{-1})$$

By the Central Limit Theorem, (19) and (20), $\hat{\theta}_n$ is $AN(\mu_n, \sigma_n/\sqrt{n})$ where σ^2 is defined as

$$(21) \qquad \sigma_n^2 = n^{-1} \sum_{i=1}^{n} E[\vartheta_1(u_i;\theta_0)]^2 .$$

Consequently, the required approximation for the variance term of MSE (11) is

$$(22) \qquad E(\hat{\theta}_n - \mu_n)^2 \doteq n^{-1}\sigma_n^2$$

Once again, if we assume n is a multiple of k and $\Pi_{nk}=\Pi_k$ for each n then,

$$(23) \qquad \sigma_n^2 = \sum_{k=1}^{k} w_k^2(\theta_0)\{\Pi_k^*(\theta^*)[1-\Pi_k^*(\theta^*)] + [\Pi_k^*(\theta^*) - \Pi_k(\theta_0)]^2\}\Pi_k/H_1(\theta_0)]^2 .$$

g) Relationship to Fisher's Information

Let the likelihood for each item be denoted by $L_i(\theta)=P_i(\theta)^{u_i}Q_i(\theta)^{1-u_i}$, then by the definition of $h(\theta,u_i)$ it is possible to show that

$$(24) \qquad h(\theta,u_i) = (\partial/\partial\theta) \log [L_i(\theta)]$$

$$(25) \qquad H(\theta) = E(\partial/\partial\theta) \log [L(\theta)]$$

$$(26) \qquad H_1(\theta) = (\partial/\partial\theta) E(\partial/\partial\theta) \log [L(\theta)]$$

$$= E(\partial^2/\partial\theta^2) \log [L(\theta)]$$

Thus, $H_1(\theta)$ is the Fisher information.

When the true and induced likelihoods agree,

$$(27) \qquad E(\partial^2/\partial\theta^2) \log L(\theta) = E[(\partial/\partial\theta) \log L(\theta)]^2.$$

Hence, $H_1(\theta)=E[h(\theta,u)]^2$ and by the definition of $\theta_1(u;\theta_0)$ in (17)

$$(28) \qquad \begin{aligned} \sigma^2 &= E[\theta_1(u;\theta_0)]^2 \\ &= E[h(\theta,u)]^2/[H_1(\theta)]^2 \\ &= 1/H_1(\theta). \end{aligned}$$

In general, when the true and induced likelihoods disagree, the reciprocal of the Fisher information of the induced likelihood does not yield the asymptotic variance of $\sqrt{n}(\hat{\theta}_n-u_n)$. However, the formula for σ^2 (23) is valid and its reciprocal could serve as a meaningful replacement for Fisher's information.

h) Derivation of the Taylor Expansion for $\hat{\theta}_n$

To show that $\hat{\theta}_n$ has the Taylor expansion given in (16) we show that it is a von Mises differentiable statistical functional. This means that there is a notion of an empirical distribution function of the observations, say F_n; and for any distribution function F that F_n may assume, there exists a functional $F\mapsto\theta(F)$ such that $\hat{\theta}_n=\theta(F_n)$ (Serfling, 1980, chapter 6).

For simplicity, let us motivate the functional idea by assuming that each item response function $P_i(\theta)$ is identical and equal to a common item response curve $\Pi(\theta)$. This is not to be confused with $E(u_i)$, although let us assume also that it does not depend on i and $E(u_i)=\Pi^*(\theta^*)$.

Next we define F_n as the usual empirical distribution function of $u_1,u_2,\ldots,u_n$. Furthermore, the estimating function (7) that determines the mle $\hat{\theta}_n$ can be rewritten as

$$(29) \qquad \sum_{i=1}^{n} w(\theta)[u_i-P(\theta)] = n\int w(\theta)[u-P(\theta)]dF_n(u) \ .$$

Thus, the desired functional $F\to\theta(F)$ is the solution to the equation

$$(30) \qquad \int w(\theta(F))[u-P(\theta(F))]dF(u)=0$$

such that $\hat{\theta}_n=\theta(F_n)$.

When $P_i(\theta)$ and $E(u_i)$ are possibly different for each item, the structure one needs to define an empirical distribution function F_n is complicated. The reason for this complexity is best described by a comparison to the identical item situation we assumed above. While the sample space for the identical item case is simply the two point set $\{0,1\}$, the general sample space is the product set $\{0,1\}\times C$ where C is the set of all pairs of item response functions $\Pi_k(\cdot)$ and $\Pi_k^*(\cdot)$. Jones (1982) developes the necessary topological machinery for this sample space in order that $\hat{\theta}_n$ is a von Mises differentiable statistical functional.

We now explain the Taylor expansion (16). Let $F_\lambda=F+\lambda(F_n-F)$ where F is some other distribution function and $0<\lambda<1$. Define $Q(\lambda)=\theta(F_\lambda)$, and expand $Q(\lambda)$ as function of λ in a standard Taylor series. To justify this expansion we require that $Q(\lambda)$ has a finite mth derivative $Q^{(m)}$ everywhere in the open interval (0,1); the kth order right-hand derivative $Q_+^{(k)}(\lambda)$ is continuous at $\lambda=0$; and the kth order left-hand derivative $Q_-^{(k)}(\lambda)$ is continuous at $\lambda=1$ for

$k=1,\ldots,m-1$ (Apostol, 1957, p.96; Serfling, 1980, pp. 43-44). Since $Q(0)=\theta(F)$ and $Q(1)=\theta(F_n)$, the Taylor expansion for $Q(\cdot)$ may be expressed as a Taylor expansion for the functional $\theta(\cdot)$:

$$(31)\quad \theta(F_n)-\theta(F) = \sum_{k=1}^{m-1} \frac{1}{k!} d_k\theta(F;F_n-F) + \frac{1}{m!}\frac{d^m}{d\lambda^m}\theta(F+\lambda(F_n-F))\Big|_{\lambda=\lambda^*}$$

where $0<\lambda^*<1$ and

$$(32)\quad d_k\theta(F;F_n-F) = (d^k/d\lambda^k)\theta(F+\lambda(F_n-F))\Big|_{\lambda=0}$$

$$= Q_+^{(k)}(0)$$

Expression (32) is called the kth Gâteau derivative of $\theta(F)$ at F in the direction F_n.

We must prove that $Q(\lambda)$ satisfies the required conditions on its derivatives. These derivatives are obtained implicitly from expression (30), thus the derivatives exist if and only if the derivatives of the integrand, $h(\theta,u)=w(\theta)[u-P(\theta)]$, with respect to θ, exist; which they do for all orders, since the derivatives of $P_1(\theta)$ exist for all orders for each response models (2), (3), and (4).

Now we formulate the derivatives $d_k\theta(F;F_n-F)$, $k=1,2$. Letting

$$(33)\quad H(\theta,\lambda) = \int h(\theta,u)dF_\lambda(u)$$

and since $H(\theta(F_\lambda),\lambda)=0$, for all $0<\lambda<1$, we have

$$(34)\quad H_1(\theta(F_\lambda),\lambda)\ (d/d\lambda)\theta(F_\lambda) + \int h(\theta,u)d(F_n-F) = 0\ .$$

Hence,

$$(35)\quad d_1\theta(F;F_n-F) = (d/d\lambda)\theta(F_\lambda)\big|_{\lambda=0} = -\frac{\int h(\theta(F),u)\,d(F_n(u)-F(u))}{H_1(\theta(F))}$$

Since $\int h(\theta(F),u)\,dF=0$,

$$(36)\quad d_1(\theta;F_n-F) = -\frac{n^{-1}\sum_{i=1}^{n} h(\theta,u_i)}{H_1(\theta(F))}$$

$$= n^{-1}\sum_{i=1}^{n}\theta_1(u_i;\theta(F))$$

The second derivative $d_2\theta(F;F_n-F)$ can be shown to be (Serfling, 1980, p.241)

$$(37)\quad d_2\theta(F;F_n-F) = \int(d/d\lambda)\theta_1(u;\theta(F_\lambda))\big|_{\lambda=0} d(F_n(u)-F(u)) .$$

Using (29) and some tedious applications of chain rule differentiation we find

$$(38)\quad (d/d\lambda)\theta_1(U;\theta(F_\lambda)\big|_{\lambda=0} = \frac{(d/d\theta)[\int h(\theta,u)h(\theta_1 v)d(F_n(v)-F(v))/H_1(\theta)]\big|_{\theta=\theta(F)}}{H_1(\theta(F))}$$

$$= \frac{n^{-1}\sum_{j=1}^{n}(d/d\theta)[h(\theta,u)h(\theta,u_j)/H_1(\theta(F))]\big|_{\theta=\theta(F)}}{H_1(\theta(F))}$$

Hence

$$= n^{-1}\sum_{j=1}^{n}\theta_2(u,u_j;\theta(F))$$

$$(39)\quad d_2\theta(F;F_n-F) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\theta_2(u_i,u_j;\theta(F))$$

i) Determination of Ψ_n Estimates

Let $\{\theta_j^*: j=1,\ldots,100\}$ represent the true abilities of 100 respondents, then the log-likelihood for estimating $\{\theta_j^*\}$ and Ψ_n is

$$(40) \qquad \sum_{j=1}^{100} \log L(\theta_j; u_{1j},\ldots,u_{nj},\Psi_n)$$

$$= \sum_{j=1}^{100} \sum_{i=1}^{n} [u_{ij} \log P_i(\theta_j,\Psi_n) + (1-u_{ij}) \log Q_i(\theta_j,\Psi_n)] .$$

Since we know the true models, we can sample indefinitely from them to simulate each respondent's behavior. Hence, a simulation would be essentially an approximation to what would have occurred if we had taken n to infinity. We need not do a simulation to determine Ψ_n since we know the limiting form of the likelihoods. As n goes to infinity, by the strong law of large numbers

$$(41) \quad n^{-1} \sum_{j=1}^{100} \log L(\theta_j; u_{1j},\ldots,u_{nj},\Psi_n) \to$$

$$\sum_{j=1}^{100} \sum_{k=1}^{k} [\Pi_k^*(\theta_j^*) \log \Pi_k(\theta_{0j};\Psi_k) + [1-\Pi_k^*(\theta_j^*)] \log [1-\Pi_k(\theta_{0j};\Psi_k)]]\Pi_k$$

Where Ψ_k denotes the parameters associated with Π_k.

We take $\Pi_k=1/k$ and minimize expression (41) for $\hat{\Psi}_n$ when n=k. Since $\Pi_k^* \neq \Pi_k$, the corresponding item parameters do not agree either. This estimate $\hat{\Psi}_n$ depends on the chosen $\{\theta_j^*\}$. It is the variability of $\hat{\Psi}_n$ arising from the sampling variability of $\{\theta_j^*\}$ that we address next. We call a set of $\{\theta_j^*\}$ the calibration sample.

j) Accounting for the Inaccuracy in Calibrated Ψ_n Estimators

We now turn to the reality that Ψ_n is a calibrated version of item parameters and hence subject to sampling variation. We note sampling variation of $\hat{\Psi}_n$ for the calibration by

$$\tilde{\Psi}_n = \hat{\Psi}_n - E_n \tag{42}$$

where $\tilde{\Psi}_n$ denotes the "true" item parameters and E_n a matrix of errors. In our situation we avoid assuming that the models for $\{E(u_i), i=1,2,\ldots n\}$ is any of the three logistic response models (2)-(4) for the calibration sample, thus the true item parameters do not correspond to any of those given in the logistic models. Therefore, we take Ψ_n to be the expected value of the asymptotic distribution of $\tilde{\Psi}_n$ as the number of individuals becomes infinite; each individual's ability θ being sampled from $N(0,1)$ during the calibration. In this asymptotic scheme, we also assume that the number of items remains fixed. The asymptotic distribution of an appropriate normed version of Ψ_n is multivariate normal and can be derived using the program of proof in Huber (1967). Moreover, Thissen and Wainer (1982) have derived appropriate approximations for the covariance matrix which we will use.

The distribution of E_n, then, is mulitivariate normal mean matrix zero and covariance matrix $\sum_n$. We will discuss the form of $\sum_n$ later, for now we will develop measures for the effect of the sampling variability of Ψ_n on the

mean squared error for a fixed value of θ^*. Define $S(\Psi_n)$ to be the set of Gaussian quadrature points centered at Ψ_n. We take this set of points to be plausible alternatives to Ψ_n. That is we view $\tilde{\Psi}_n$, the "true" parameters, as lying in $S(\Psi_n)$.

Denote the quadrature points by

$$\{\tilde{\Psi}_{1n}, \tilde{\Psi}_{2n}, \ldots \tilde{\Psi}_{Jn}\} = S(\Psi n) \tag{43}$$

with corresponding Gaussian quadrature weights

$$\{\lambda_{1n}, \lambda_{2n}, \ldots, \lambda_{Jn}\} \ . \tag{44}$$

We will assume that the weights are normed so that

$$\sum_{j=1}^{J} \lambda_{jn} = 1.$$

Denote the mean squared errors for each point in $S(\Psi_n)$ by

$$MSE_j = MSE(\tilde{\Psi}_{jn}, \theta^*) \ . \tag{45}$$

Further denote the ordered values of these MSE's and their associated weights by

$$MSE_{(1)} < MSE_{(2)} < \ldots < MSE_{(J)} \text{ and} \tag{46}$$

$$\lambda_{(1)}, \lambda_{(2)}, \ldots, \lambda_{(J)} \text{ respectively.} \tag{47}$$

We may view $MSE(\Psi_n,\theta^*)$ as a positive random variable for each θ^* where we have approximated its distribution by

$$(48) \qquad Pr[MSE(\Psi n,\theta^*) = MSE_j] = \lambda_j \ .$$

We characterize this distribution by the mean, the maximum, the minimum and percentiles as follows. The mean is the weighted average denoted as the unconditional MSE

$$(49) \qquad UNCOND.\ MSE = \sum_{j=1}^{J} MSE_j \lambda_j \ .$$

The cumulative distribution function is approximated by

$$(50) \qquad F_n(t) = \sum_{j \in A_t} \lambda_{(j)}$$

where $A_t = \{j:MSE_{(j)} < t\}$ and hence the p x 100-th percentile is

$$(51) \qquad P.MSE = F_n^{-1}(p) \ .$$

We give special designations to the 25-th and 75-th percentiles denoted by

$$(52) \qquad Q1.MSE = F_n^{-1}(.25)$$

$$(53) \qquad Q3.MSE = F_n^{-1}(.75)$$

The minimum and maximum is denoted by

$$\text{MIN.MSE} = \text{MSE}_{(1)} = F_n^{-1}(0) \tag{54}$$

$$\text{MAX.MSE} = \text{MSE}_{(J)} = F_n^{-1}(1) \ . \tag{55}$$

We will compare the unconditional MSE (49) to the value $\text{MSE}(\Psi_n,\theta^*)$ and for this purpose we call the latter value the conditional MSE and give it the special notation

$$\text{COND MSE} = \text{MSE}(\Psi_n,\theta^*) \ . \tag{56}$$

The covariance matrix of $\sum_n$ is formulated for our problem as follows. Let elements of E be denoted as

$$E = \{e_{ij}: i=1,2,\ldots,n;\ j=1,2,3\} \tag{57}$$

and let the variance e_{ij} be denoted by σ_{ij}^2. Then

$$\sum_n = \{\text{cov}(e_{ij},e_{i'j'}): i,i'=1,2,\ldots n;\ j,j'=1,2,3\} \tag{58}$$

where

$$\text{cov}(e_{ij},e_{i'j'}) = \begin{cases} \sigma_{ij}^2 \text{ for } i=i',\ j=j' \\ -\sigma_{ij}\,\sigma_{i'j'} \text{ for } i=i',\ j\neq j' \\ 0 \text{ for } i\neq i'. \end{cases}$$

The above covariance structure may be written more compactly and hence more transparently as follows. Denote by $\text{DIAG}(D_1,D_2,\ldots,D_n)$ the matrix that has the matrices $\{D_i,i=1,\ldots,n\}$ down its diagonal. For instance, if each D_i

is a 3x3 matrix then DIAG($D_1, D_2, \ldots, D_n$) is a $3_n x 3_n$ matrix. Using this notation, the above Σ_n in (33) may be written as Σ_n = DIAG($D_1, D_2, \ldots, D_n$) where

$$D_i = \begin{bmatrix} \sigma_{i1}^2 - \sigma_{i1}\sigma_{i2} - \sigma_{i1}\sigma_{i3} \\ \sigma_{i2}^2 - \sigma_{i2}\sigma_{i3} \\ \text{(symmetric)} \quad \sigma_{i3}^2 \end{bmatrix}$$

as i=1,2,...,n.

TABLE 1

Probability of Correct Response for Four ASVAB Items

	TRUE ABILITY				
Item	-2	-1	0	1	2
1	.42	.32	.19	.50	.89
2	.27	.26	.28	.44	.77
3	.03	.05	.11	.47	.88
4	.36	.26	.20	.60	.95

TABLE 2

Estimates of Item Characteristic Parameters of Four ASVAB Items (s.e.)

1-PL

Item	DIFFICULTY	DISCRIMINATION	LOWER ASYMPTOTE
1	.9 (.3)	.7 (.1)	0
2	1.0 (.3)	.7 (.1)	0
3	1.9 (.3)	.7 (.1)	0
4	.9 (.3)	.7 (.1)	0

2-PL

Item	DIFFICULTY	DISCRIMINATION	LOWER ASYMPTOTE
1	1.4 (.6)	.5 (.2)	0
2	1.4 (.6)	.6 (.2)	0
3	1.1 (.2)	1.7 (.3)	0
4	.9 (.4)	.7 (.2)	0

3-PL

Item	DIFFICULTY	DISCRIMINATION	LOWER ASYMPTOTE
1	1.3 (.2)	3.2 (1.5)	.26 (.08)
2	1.6 (.6)	1.9 (1.9)	.25 (.07)
3	1.1 (.2)	2.1 (.5)	.03 (.02)
4	1.1 (.2)	3.5 (1.4)	.23 (.07)

TABLE 3

Asymptotic Values of Error and Its Components
Evaluated at the Point of the Estimate
for n=4

True Ability	θ_0: Limit of $\hat{\theta}_n$			$n\times(\text{Bias})^2$			σ_n^2 = n x Variance $(\hat{\theta})$			n x (M.S.E.)		
	1-PL	2-PL	3-PL	1-PL	2-PL	3-PL	1-PL	2-PL	3-PL	1-PL	2-PL	3-PL
-2	-.2	-.3	-1.2	12.9	11.2	2.3	9.2	7.2	617.5	22.1	18.4	619.8
-1	-.6	-.6	-.8	.7	.7	.2	10.7	11.0	114.5	11.4	11.6	114.7
0	-.9	-.6	-.1	2.7	1.3	.0	12.2	14.9	18.5	14.8	16.2	18.5
1	1.2	1.1	1.0	.2	.0	.0	7.9	4.2	1.0	8.1	4.2	1.0
2	3.8	3.3	1.9	13.2	6.4	.0	17.0	33.4	1.7	30.3	39.7	1.8

TABLE 4

A Comparison of the M.S.E. of Ability at a Point
(Conditioned Estimates) and Integrated Over a Region
(Unconditioned Estimates)

	Ability									
	-2		-1		0		1		2	
	C.	Unc.	C.	Unc.	C.	Unc.	C.	Unc.	C.	Unc.
1-PL	22	22	11	12	15	15	8	8	30	31
2-PL	18	*	12	19	16	20	4	5	40	41
3-PL	619	*	115	*	19	*	1	2	2	2

*Indicates m.s.e. larger than 10^4.

TABLE 5

M.S.E. in the Region of Conditional Estimate

	-2			-1			0			1			2		
	1-PL	2-PL	3-PL	1-PL	2-PL	3-PL	1-PL	2-PL	3-PL	1-PL	2-PL	3-PL	1-PL	2-PL	3-PL
Min.	18	18	19	11	9	12	13	6	18	8	3	1	27	26	1
25%ile	22	20	24	11	10	23	14	10	21	8	4	1	27	34	2
75%ile	24	*	*	12	41	69	15	17	47	8	6	2	32	45	2
Max.	27	*	*	13	80	*	17	66	*	9	8	3	35	96	5
Cond.	22	18	619	11	12	115	15	16	19	8	4	1	30	40	2

* = Larger than 10^4

TABLE 6

Relative Efficiency in the Region of Conditional Estimate

	-2			-1			0			1			2		
MODEL	1-PL	2-PL	3-PL	1-PL	2-PL	3-PL	1-PL	2-PL	3-PL	1-PL	2-PL	3-PL	1-PL	2-PL	3-PL
Min.	67	0	0	71	11	0	37	10	0	10	12	36	3	1	24
25%ile	75	0	0	74	50	11	40	32	13	10	15	59	4	3	62
75%ile	89	91	67	80	86	39	43	64	31	11	25	96	4	4	79
Max.	100	100	92	84	100	67	48	100	36	12	27	100	5	5	100
Cond.	81	98	3	79	77	8	43	39	34	10	22	98	4	3	70

TABLE 7

Asymptotic Values of Error and Its Components
Evaluated at the Point of the Estimate

True Ability	Expected Value			$n(\text{Bias})^2$			$n \times$ variance $(\hat{\theta}_n) \doteq \sigma_0^2$			$n \times$ (M.S.E.)			n	Variance $(\hat{\theta}_n) = \sigma_0^2/n$			M.S.E.		
	1-PL	2-PL	3-PL	1-PL	2-PL	3-PL	1-PL	2-PL	3-PL	1-PL	2-PL	3-PL		1-PL	2-PL	3-PL	1-PL	2-PL	3-PL
-2	-.2	-.3	-1.2	12.9	11.2	2.3	9	7	618	22	18	620		2.3	1.8	154.4	5.5	4.6	154.9
-1	-.6	-.6	-.8	.7	.7	.2	11	11	115	11	11	115		2.7	2.8	28.6	2.9	2.9	28.7
0	-.9	-.6	-.1	2.7	1.3	0	12	15	19	15	16	19	4	3.1	3.7	4.6	3.7	4.1	4.6
1	1.2	1.1	1.0	.2	0	0	8	4	1	8	4	1		2.0	1.1	.3	2.0	1.1	.3
2	3.8	3.3	1.9	13.2	6.4	0	17	33	2	30	40	2		4.3	8.4	.4	7.6	9.9	.5
-2				51.6	44.8	9.2				61	54	627		.6	.5	38.6	3.8	3.4	39.2
-1				2.8	2.8	.8				14	14	115		.7	.7	7.2	.8	.9	7.2
0	Same as Above			10.8	5.2	0	Same as Above			23	20	19	16	.8	.9	1.2	1.4	1.3	1.2
1				.8	0	0				9	4	1		.5	.3	.1	.5	.3	.1
2				52.8	25.6	0				70	59	2		1.1	2.1	.1	4.4	3.7	.1
-2				154.8	134.4	27.6				164	142	645		.2	.2	12.9	3.4	3.0	13.4
-1				8.4	8.4	2.4				19	19	117		.2	.2	2.4	.4	.4	2.4
0	Same as Above			32.4	15.6	0	Same as Above			45	31	19	48	.3	.3	.4	.9	.6	.4
1				2.4	0	0				10	4	1		.2	.1	0	.2	.1	0
2				158.4	76.8	0				175	110	2		.4	.7	0	3.7	2.3	0
-2				309.6	268.8	55.2				319	276	673		.1	.1	6.4	3.3	2.9	7.0
-1				16.8	16.8	4.8				28	28	119		.1	.1	1.2	.3	.3	1.2
0	Same as Above			64.8	31.2	0	Same as Above			77	46	19	96	.1	.2	.2	.8	.5	.2
1				4.8	0	0				13	4	1		.1	0	0	.1	0	0
2				316.8	153.6	0				334	187	2		.2	.3	0	3.5	1.9	0
														0	0	0	3.2	2.8	.6
														0	0	0	.2	.2	.1
													∞	0	0	0	.7	.3	0
														0	0	0	.1	0	0
														0	0	0	3.3	1.6	0

Figure Captions

Figure 1 - Item characteristic curves fitted to four ASVAB items using the BSBT Model.

Figure 2 - Item characteristic curves fitted to four ASVAB items using the 1-PL Model (dashed line) shown with the BSBT curves (solid line).

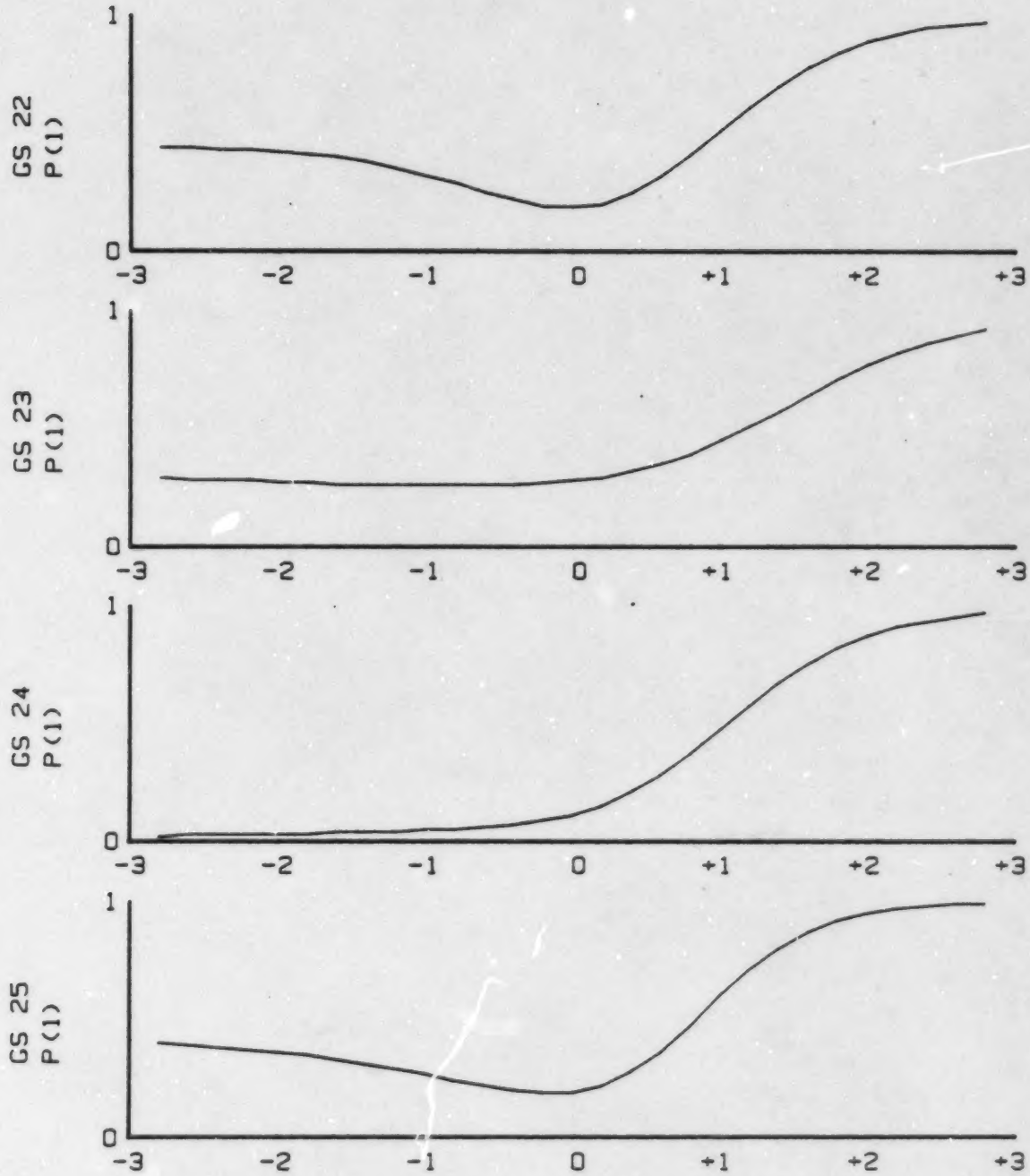
Figure 3 - Item characteristic curves fitted to four ASVAB items using the 2-PL Model (dashed line) shown with the BSBT curves (solid line).

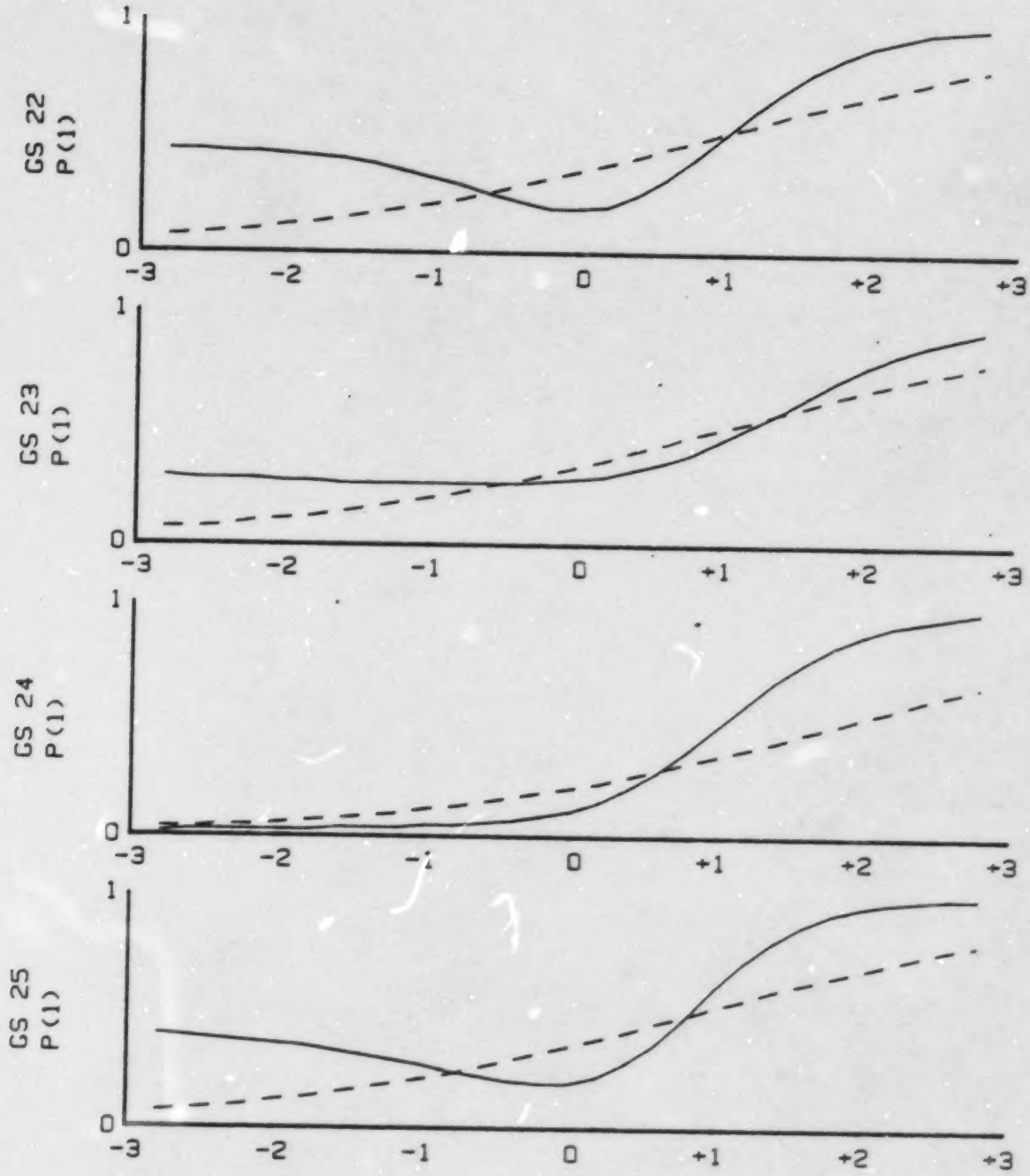
Figure 4 - Item characteristic curves fitted to four ASVAB items using the 3-PL Model (dashed line) shown with the BSBT curves (solid line).

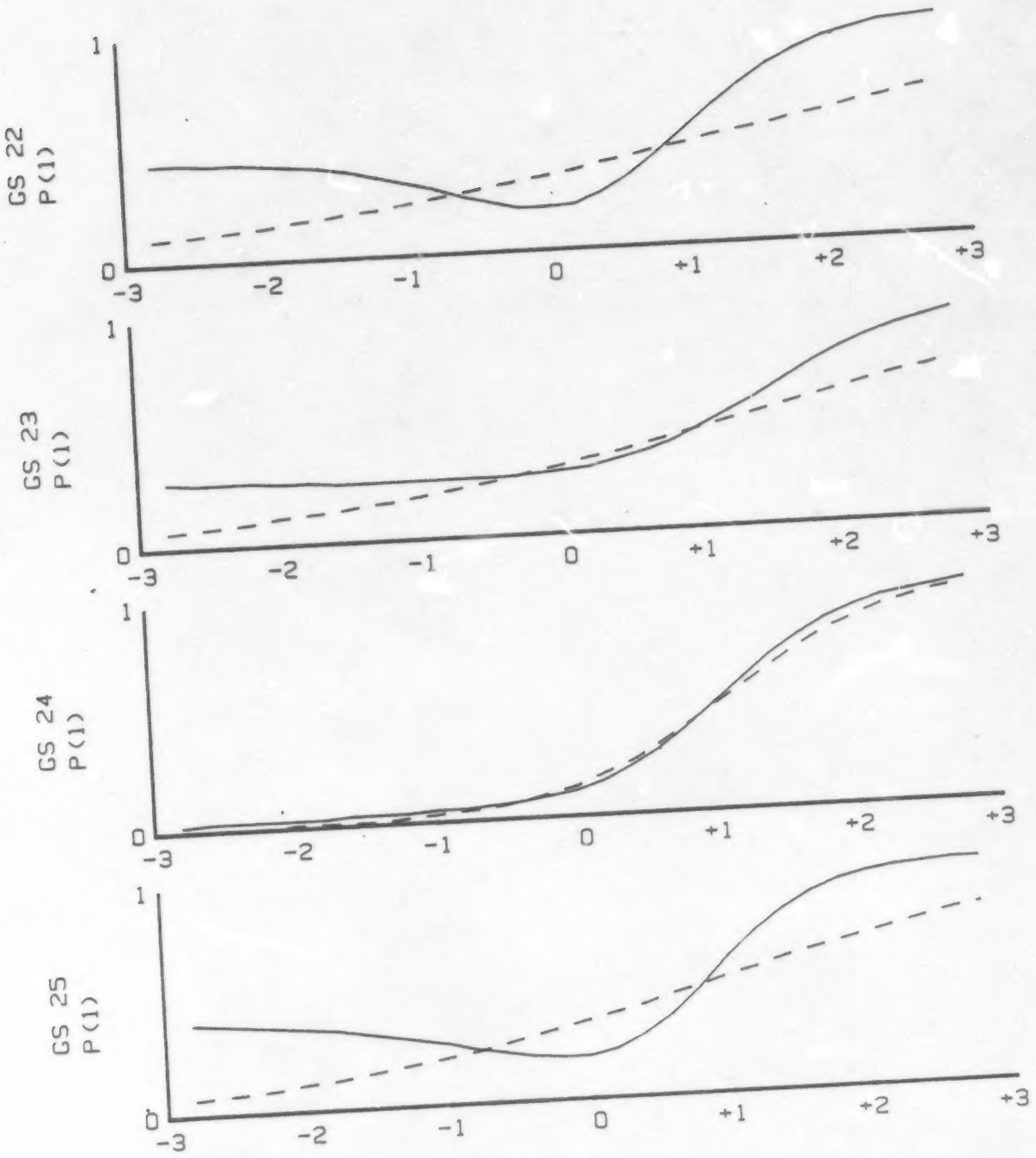
Figure 5 - Relative information loss for the three logistic models that occurs when one uses the unconditional estimate rather than the conditional estimate.

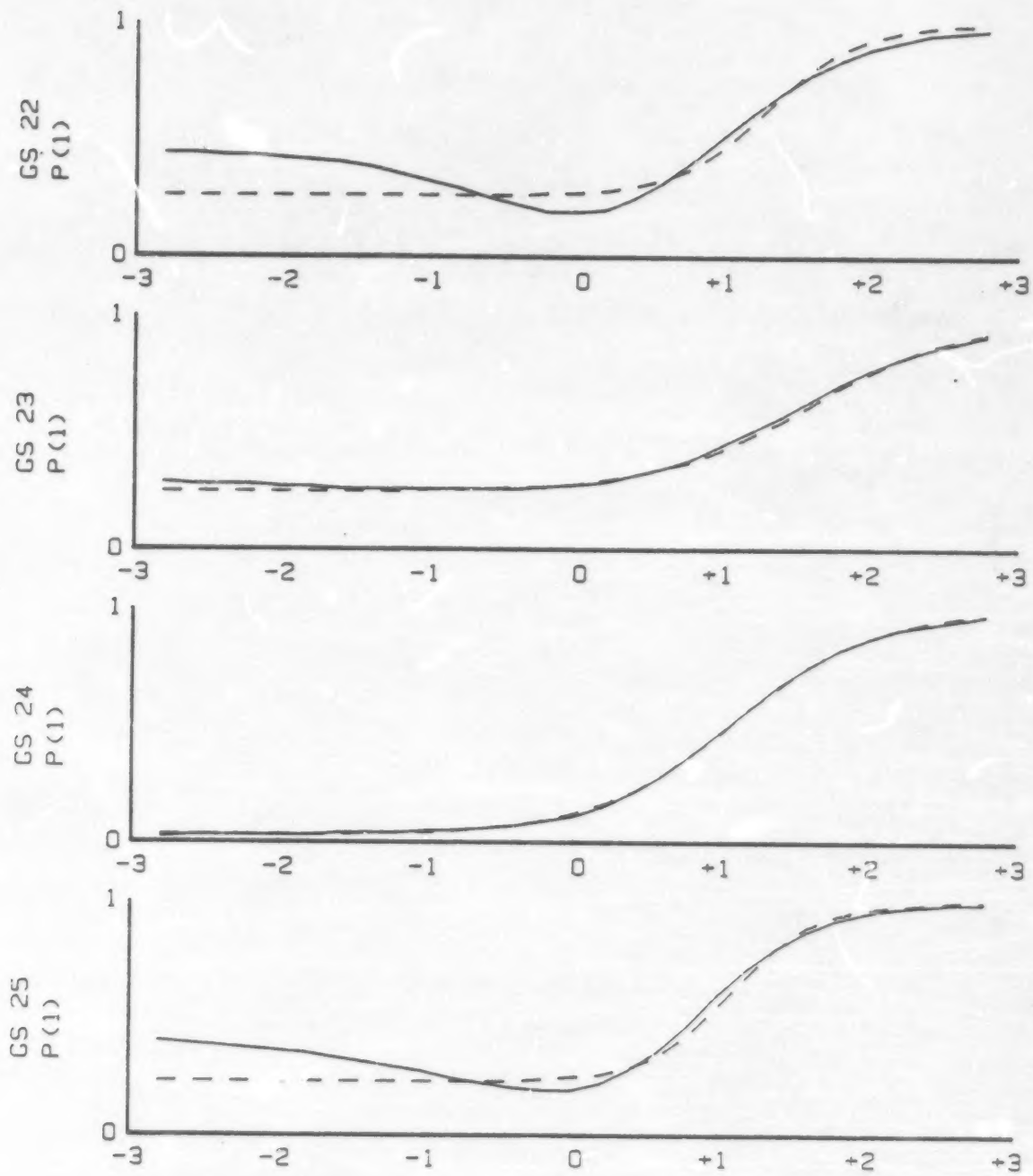
Figure 6 - The proportion of the mean square error that is caused by bias in the estimator of ability for each of the three models.

Figure 7 - Empirical information functions for the mle of ability in each of the three models.


GS 22
P(1)
1
0
-3
-2
-1
0
+1
+2
+3
GS 23
P(1)
1
0
-3
-2
-1
0
+1
+2
+3
GS 24
P(1)
1
0
-3
-2
-1
0
+1
+2
+3
GS 25
P(1)
1
0
-3
-2
-1
0
+1
+2
+3

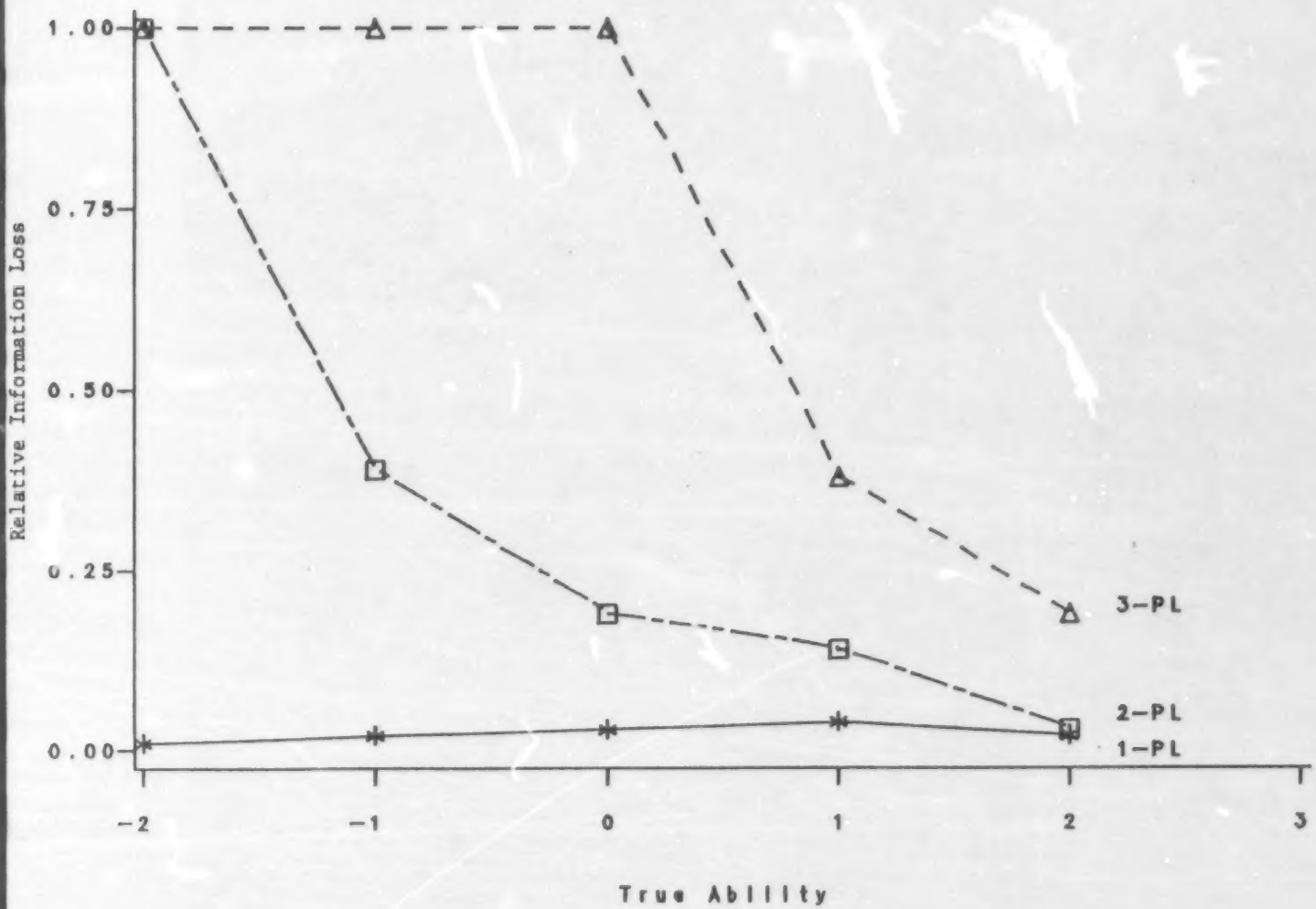

GS 22
P(1)
1
0
-3
-2
-1
0
+1
+2
+3
GS 23
P(1)
1
0
-3
-2
-1
0
+1
+2
+3
GS 24
P(1)
1
0
-3
-2
-1
0
+1
+2
+3
GS 25
P(1)
1
0
-3
-2
-1
0
+1
+2
+3


GS 22
P(1)
1
0
-3
-2
-1
0
+1
+2
+3
GS 23
P(1)
1
0
-3
-2
-1
0
+1
+2
+3
GS 24
P(1)
1
0
-3
-2
-1
0
+1
+2
+3
GS 25
P(1)
1
0
-3
-2
-1
0
+1
+2
+3

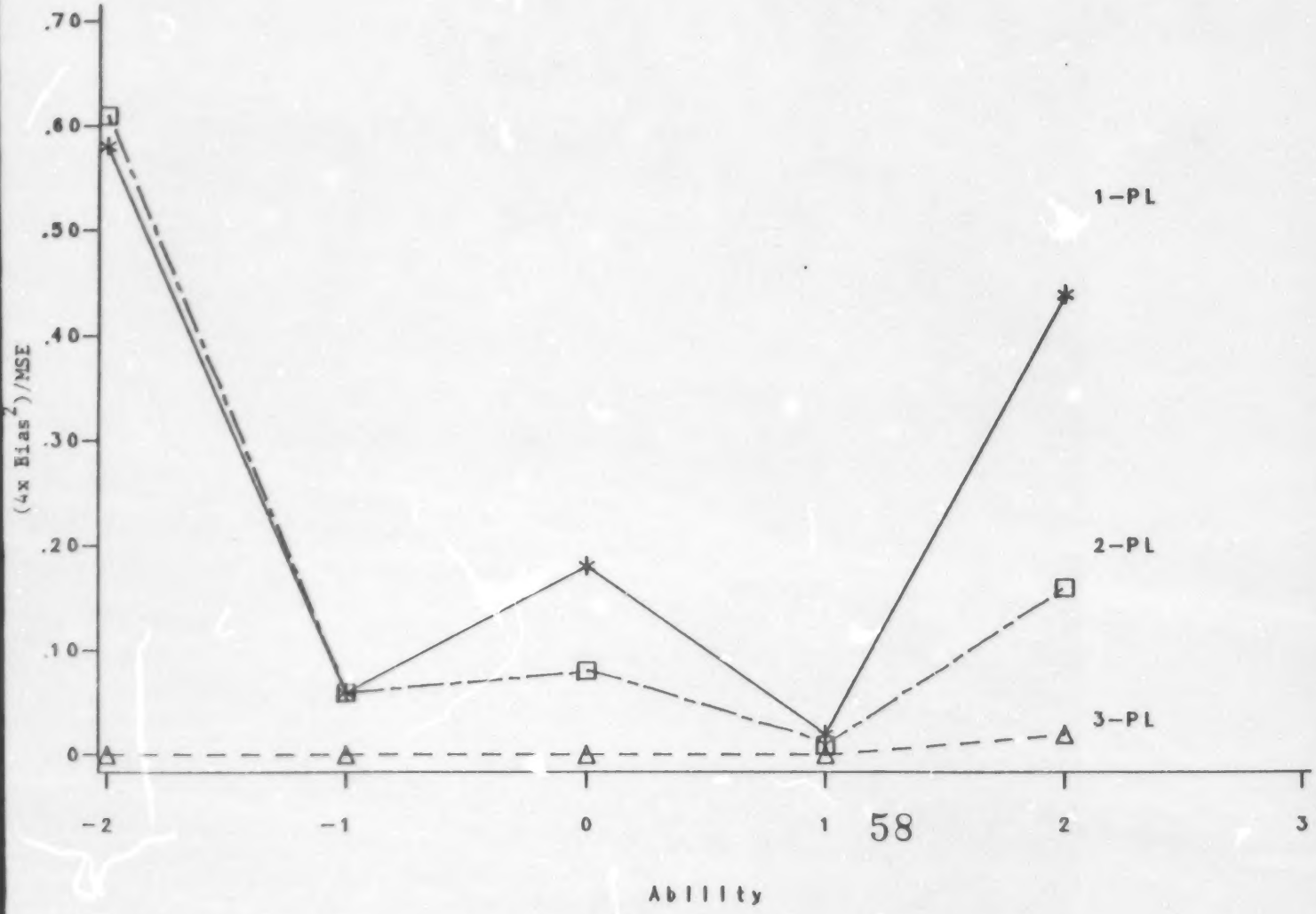

GS 22
P(1)
1
0
-3
-2
-1
0
+1
+2
+3
GS 23
P(1)
1
0
-3
-2
-1
0
+1
+2
+3
GS 24
P(1)
1
0
-3
-2
-1
0
+1
+2
+3
GS 25
P(1)
1
0
-3
-2
-1
0
+1
+2
+3

Relative Information Loss

From Conditional to Unconditional MSE



Proportion of Bias Squared in MSE



Reciprocal of Conditional MSE,

The information in an ability estimator